Method for running with a Meta search engine adapting to a new search response display processes a search response delivered by a primary search engine in a search response display

Patent number:	DE10056681
Publication date:	2002-05-23
Inventor:	KRUG ADRIAN (DE); MOLL CHRISTOPH (DE)
Applicant:	HEWLETT PACKARD CO (US)
Classification:	
- international:	G06F15/18; G06F17/30; G06F15/177
- european:	G06F17/30W1
Application number:	DE20001056681 20001115
Priority number(s):	DE20001056681 20001115

Abstract of **DE10056681**

A Meta search engine server (4) acts as an interface between a user host computer (2) and multiple primary search engine (6) servers. Instead of sending a separate search query to all servers for the primary search engines (PSE), the user host computer directs its inquiry only once at the Meta search engine server that adapts the inquiry to the special requirements of the PSEs and transmits the special search queries to the individual servers of the PSEs. Independent claims are also included for (1) a computer system with a Meta search engine with an interface to a primary search engine (2) a computer program product with program code for running the method of the present invention.

(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENT- UND MARKENAMT

(12) **Offenlegungsschrift**

(10) **DE 100 56 681 A 1**

(51) Int. Cl.[7]:
G 06 F 15/18
G 06 F 17/30
G 06 F 15/177

(21) Aktenzeichen: 100 56 681.2
(22) Anmeldetag: 15. 11. 2000
(43) Offenlegungstag: 23. 5. 2002

(71) Anmelder:
Hewlett-Packard Co. (n.d.Ges.d.Staates Delaware), Palo Alto, Calif., US

(74) Vertreter:
Samson & Partner, Patentanwälte, 80538 München

(72) Erfinder:
Krug, Adrian, 40595 Düsseldorf, DE; Moll, Christoph, 46238 Bottrop, DE

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

Prüfungsantrag gem. § 44 PatG ist gestellt

(54) Verfahren, Computersystem und Computerprogramm-Produkt zum Konfigurieren einer Meta-Suchmaschine

(57) Die Erfindung ist auf ein von einer Meta-Suchmaschine durchgeführtes Verfahren gerichtet. In dem Verfahren wird eine Suchantwort, die in einer Suchantwortdarstellung von der Primärsuchmaschine geliefert wird, von der Meta-Suchmaschine verarbeitet. Das Verfahren umfaßt, daß die Meta-Suchmaschine sich selbst an eine neue Suchantwortdarstellung anpaßt. Die Erfindung ist auch auf ein von einem Computersystem durchgeführtes Verfahren gerichtet, um eine Schnittstelle zu mindestens einer Primärsuchmaschine zu konfigurieren. Die Schnittstelle hat die Funktion, Suchergebnisse aus einer Suchantwort einer Primärsuchmaschine in einer Suchantwortdarstellung zu extrahieren. Das Verfahren umfaßt das automatische Anpassen der Schnittstelle an eine neue Suchantwortdarstellung. Die Erfindung ist auch auf ein entsprechendes Computersystem und ein entsprechendes Computerprogramm-Produkt gerichtet.

DE 100 56 681 A 1

DE 100 56 681 A 1

Beschreibung

[0001] Die vorliegende Erfindung betrifft allgemein Meta-Suchmaschinen, und genauer Verfahren, ein Computersystem und ein Computerprogramm-Produkt zum Konfigurieren einer Meta-Suchmaschine, so daß Suchantworten von Primärsuchmaschinen verarbeitet werden.

[0002] Die Menge an Informationen, die über Netzwerke und Online-Datenbanken zur Verfügung steht, hat rasch zugenommen und nimmt weiter zu. Besonders der weit verbreitetste Dienst im Internet, das World Wide Web (WWW), hat in den letzten 5 Jahren einen explosionsartigen Wachstum erlebt. Andererseits wird das Ausfindigmachen von Informationen im Internet immer schwieriger. Getrieben durch seine offene und unkontrollierte Organisationsstruktur, werden die Informationen unstrukturiert gespeichert und machen es so dem Benutzer schwer, an Informationen zu einem speziellen Thema zu gelangen. Es gibt insbesondere kein zentrales Archiv, das als Verweis zu Informationen im Internet dient. Des weiteren kann keine Filterung oder irgendeine andere Kontrolle der Informationen angewandt werden, um die Zugänglichkeit der im World Wide Web verfügbaren Dokumente zu verbessern. Sogar innerhalb einer einzigen Web-Site ist es für den Benutzer oft schwer, nur durch Navigieren entlang der bereitgestellten Hyperlinks (Verweise zu WWW-Dokumenten), die gewünschten Informationen zu finden. Darüber hinaus bieten immer mehr Firmen ihren Kunden und Angestellten einen zusätzlichen Service in Form von umfangreichen Informationen über ihre Produkte und Dienstleistungen. Da diese Informationsdienste gewöhnlich sowohl auf das Internet als auch auf firmeninterne Netzwerke (Intranet), die auf Internettechnologien basieren, zugreifen, ist ihre Struktur der des Internets ähnlich. Außerdem hat die Menge an Informationen, die durch diese Dienste zur Verfügung gestellt wird, für Kunden und Angestellte eine handhabe Größe überschritten. Folglich herrscht eine starke Nachfrage nach Werkzeugen, die die Informationsbeschaffung im Internet, Intranet oder auf großen Web-Sites erleichtert. Werkzeuge, die in der Lage sind, im Internet oder Intranet nach spezifischen Informationen zu suchen, werden Suchmaschinen genannt.

[0003] Suchmaschinen versetzen den Benutzer in die Lage, in Webseiten nach spezifischen Stichworten zu suchen. Sie basieren in der Regel auf suchfähigen Datenbanken oder Archiven, in denen Querverweise zu Web-Sites, sogenannte Uniform Resource Locators (URL), abgelegt sind. Zusammen mit der URL werden die wichtigsten Site-Informationen gespeichert, d. h. Stichworte und Begriffe, die in dem entsprechenden Dokument enthalten sind, sowie eine kurze Beschreibung des Inhalts der Seite. Spezielle Programme, sogenannte "Spinnen" oder "Webroboter", die das Web laufend nach neuen Sites durchsuchen und Stichworte identifizieren, helfen der Suchmaschine die Datenbank zu ergänzen und zu aktualisieren.

[0004] In den letzten Jahren haben sich eine Reihe von Suchmaschinen etabliert, von denen die gängisten unter www.altavista.com, www.lycos.com, www.excite.com oder www.yahoo.com gefunden werden können. Zusätzlich spezialisieren sich viele andere Suchmaschinen auf spezielle Felder, z. B. auf Patentsuche (www.patents.ibm.com), lokale Informationen (www.bigyellow.com), Software (www.tucows.com), Jobs (www.carreerbuilder.com) oder Musik (www.scour.net). Weitere Beispiele für Suchmaschinen sind Intranetsuchmaschinen, die ihren Suchbereich auf ein internes Firmen-, Instituts-, oder Universitätsnetz begrenzen.

[0005] Suchmaschinen stellen dem Benutzer über eine Webseite eine Benutzerschnittstelle zur Verfügung, die es

dem Benutzer erlaubt, Stichworte oder logische Verknüpfungen von Stichworten einzugeben. Zum Beispiel würde eine Suchanfrage, die eine logische UND-Verknüpfung der Stichworte "Computer" und "Spiele" verwendet, alle in der Datenbank der befragten Suchmaschine enthaltenen Querverweise zu Web-Sites ergeben, die Informationen sowohl zu Computern als auch zu Spielen beinhalten. In der Regel werden die von einer Suchmaschine erhaltenen Ergebnisse einer Suchanfrage aufgelistet und im Browser des Benutzers, geordnet nach der Relevanz der Dokumente, angezeigt, wobei jedes Listenelement die URL, die kurze Beschreibung des Inhalts und das Datum des Dokuments enthält.

[0006] Im allgemeinen wünscht sich ein Benutzer, mehrere verschiedene Suchmaschinen zu benutzen, um die Verläßlichkeit der Suche zu erhöhen. Mit zunehmender Anzahl an Suchmaschinen wird er jedoch mit vielen verschiedenen Arten von Benutzerschnittstellen und Darstellungen der Suchergebnisse konfrontiert. Da jede Suchmaschine ihre eigene individuelle Benutzerschnittstelle und Optionen zum Konfigurieren und Optimieren der Suche hat, muß der Benutzer lernen, mit den verschiedenen Benutzerschnittstellen umzugehen und sich die Unterschiede zu merken. Zum Beispiel variiert zwischen den verschiedenen Suchmaschinen die Syntax, um eine logische Verknüpfung von Stichworten oder Stichworte, die aus mehreren getrennten Wörtern bestehen, einzugeben, oder die Art, wie Groß- und Kleinschreibung in einem Suchanfragetext interpretiert werden.

[0007] Zusätzlich ist es schwierig, insbesondere für den unerfahrenen Benutzer, einen Überblick über bestehende Suchmaschinenanbieter zu bewahren und den besten für ein spezielles Interessengebiet auszuwählen. Um sicherzustellen, daß er die besten verfügbaren Informationen im Netz bekommt, muß der Benutzer in der Regel mehrere Suchmaschinen konsultieren, dieselbe Suchanfrage in mehreren Web-Sites eingeben und dabei verschiedene Benutzerschnittstellen und Konfigurationen verwenden, und schließlich die Suchergebnisse der verschiedenen Suchmaschinen vergleichen, bewerten und ordnen. Firmeninterne Informationsdienste basieren darüber hinaus in der Regel auf verschiedenen Online-Datenbanken, die jede ein individuelles Suchwerkzeug erfordern. Insgesamt herrscht ein großes Bedürfnis, die verfügbaren Dienste zu bündeln, so daß der Benutzer auf sie nur über eine einzige Benutzerschnittstelle zugreifen kann.

[0008] Daher erschienen kürzlich immer mehr Meta-Suchmaschinen im World Wide Web und in firmeninternen Netzen, um die Qualität des Prozesses der Informationsbeschaffung im Internet oder Intranet zu verbessern und um die obigen Unzulänglichkeiten für den Benutzer zu beseitigen, die durch die wachsende Zahl an verfügbaren Suchdiensten entstehen. Einige der gängisten Meta-Suchmaschinen sind zum Beispiel Dogpile (www.dogpile.com), MetaCrawler (www.metacrawler.com), Mamma (www.mamma.com), Inference Find (www.inference.com), Find.de (www.find.de), ProFusion (www.profusion.com), Search4 (www.search4.com).

[0009] Eine Meta-Suchmaschine ist nicht eine "Suchmaschine" im buchstäblichen Sinne, da sie nicht eine Suche ausführt, sondern vielmehr die Funktion einer Schnittstelle zu Primärsuchmaschinen hat. Von Firmen zur Verfügung gestellte Meta-Suchmaschinen, ermöglichen dem Kunden und den Angestellten einen zentralen Einstiegspunkt, um in verschiedenen internen und externen Datenbanken nach Informationen oder Lösungen zu suchen, die in Zusammenhang mit den Produkten und Dienstleistungen der Firma stehen. Im Prinzip sendet die Meta-Suchmaschine unter Verwendung des Hypertext Transfer Protocols (HTTP) Suchanfragen gleichzeitig zu mehreren Primärsuchmaschinen und

bündelt die erhaltenen Suchergebnisse. Es gibt eine gemeinsame Benutzerschnittstelle für alle Suchmaschinen, die dazu verwendet wird, eine Suchanfrage einzugeben. Die Meta-Suchmaschine überträgt eine Anfrage weiter zu den Primärsuchmaschinen und wandelt die Anfrage inklusive spezieller Suchoptionen in die individuelle Syntax jeder Primärsuchmaschine um. In einigen Fällen kann der Benutzer seine bevorzugten Primärsuchmaschinen aus einer von der Meta-Suchmaschine zur Verfügung gestellten Liste auswählen. Die von den verschiedenen Primärsuchmaschinen zurückgegebenen Suchergebnisse, werden dann von der Meta-Suchmaschine verarbeitet, um 1) Treffer (Querverweise zu Webseiten, die während der Suche gefunden wurden) auszufiltern, die in den Suchergebnissen von mehr als einer Suchmaschine erscheinen, 2) die Treffer bezüglich einer von den Primärsuchmaschinen bereitgestellten Wertung zu klassifizieren, und 3) die Treffer in einem einheitlichen Layout anzuzeigen. Detailliertere Beschreibungen von Meta-Suchmaschinen können zum Beispiel bei www.metacrawler.com/help/fax/howworks.html oder bei www.mamma.com/about.html gefunden werden.

[0010] Eine der Aufgaben einer Meta-Suchmaschine ist es, die Suchergebnisinformationen von den Antwortseiten der Primärsuchmaschinen zu extrahieren. Nachdem die Meta-Suchmaschine eine Suchanfrage als HTTP-Anfrage zu einer Primärsuchmaschine gesendet hat, empfängt sie von ihr via HTTP die gefundenen Suchinformationen, d. h. eine in eine Antwortseite eingebettete Trefferliste. Da das Layout der Antwortseiten der Primärsuchmaschine nicht standardisiert ist, d. h. die verschiedenen Primärsuchmaschinen stellen ihre Suchergebnisse unterschiedlich am Bildschirm dar, ist die Meta-Suchmaschine so konfiguriert, daß sie mit den unterschiedlichen Layouts und Formaten der Suchergebnisse der verschiedenen Primärsuchmaschinen zurecht kommt. Des weiteren wird eine neue Konfiguration integriert, wenn eine zusätzliche Primärsuchmaschine zur Meta-Suchmaschine hinzugefügt wird. Darüberhinaus kann sich das Layout der Suchergebnisse von Zeit zu Zeit ändern. Deshalb werden die verschiedenen Konfigurationen auch regelmäßig überwacht und, wenn Änderungen auftreten, angepaßt.

[0011] Gemäß einem ersten Aspekt, wird in einem von einer Meta-Suchmaschine durchgeführten Verfahren eine Suchantwort, die von einer Primärsuchmaschine in einer Suchantwortdarstellung bereitgestellt wird, von der Meta-Suchmaschine verarbeitet. Das Verfahren umfaßt, daß sich die Meta-Suchmaschine selbst an eine neue Suchantwortdarstellung anpaßt.

[0012] Gemäß einem anderen Aspekt stellt die Erfindung ein von einem Computersystem durchgeführtes Verfahren bereit, um eine Schnittstelle zu mindestens einer Primärsuchmaschine zu konfigurieren. Die Schnittstelle hat die Funktion, Suchergebnisse aus Suchantworten der Primärsuchmaschinen in einer Suchantwortdarstellung zu extrahieren. Das Verfahren umfaßt das automatische Anpassen der Schnittstelle an eine neue Suchantwortdarstellung.

[0013] Gemäß einem weiteren Aspekt stellt die Erfindung ein Computersystem bereit, das eine Meta-Suchmaschine und eine Konfigurationseinheit umfaßt.

[0014] Die Meta-Suchmaschine umfaßt eine Schnittstelle zu mindestens einer Primärsuchmaschine. Die Konfigurationseinheit ist derart ausgestaltet, daß sie die Schnittstelle automatisch an eine neue Suchantwortdarstellung der Primärsuchmaschine anpaßt.

[0015] Gemäß noch einem weiteren Aspekt stellt die Erfindung ein Computerprogramm-Produkt mit Programmcode bereit, um ein Verfahren zum Konfigurieren einer Schnittstelle zu mindestens einer Primärsuchmaschine durchzuführen, wenn es auf einem Computersystem ausgeführt wird. Die Schnittstelle hat die Funktion, Suchergebnisse aus einer Suchantwort der Primärsuchmaschine in einer Suchantwortdarstellung zu extrahieren. Das Verfahren umfaßt die automatische Anpassung der Schnittstelle an eine neue Suchantwortdarstellung.

[0016] Andere Merkmale sind dem offenbarten Verfahren und der offenbarten Vorrichtung inherent oder werden durch die folgende detaillierte Beschreibung der Ausführungsbeispiele und der zugehörigen Zeichnungen dem Fachmann ersichtlich werden.

[0017] Im folgenden werden die Zeichnungen kurz beschrieben.

[0018] **Fig.** 1 ist ein Blockdiagramm, das die Architektur eines Systems auf hoher Ebene visualisiert, das eine Meta-Suchmaschine, eine Primärsuchmaschine und einen Benutzer-Hostcomputer umfaßt;

[0019] **Fig.** 2 ist eine funktionelle Darstellung einer Schnittstelle zwischen einer Meta-Suchmaschine und einer Primärsuchmaschine;

[0020] **Fig.** 3 ist ein Blockdiagramm, das die Extraktion von Suchergebnisinformationen veranschaulicht;

[0021] **Fig.** 4 ist ein Blockdiagramm, das die automatische Erkennung von neuen Suchantwortdarstellungen veranschaulicht;

[0022] **Fig.** 5 zeigt eine typische Suchantwort einer Primärsuchmaschine;

[0023] **Fig.** 6 stellt den HTML-Quellcode eines speziellen Suchergebnisrahmens dar;

[0024] **Fig.** 7 stellt den zu dem Suchergebnisrahmen von **Fig.** 6 gehörenden HTML-Syntaxbaum dar;

[0025] **Fig.** 8 stellt einen dreidimensionalen Merkmalsraum für HTML-Syntaxelemente dar;

[0026] **Fig.** 9 zeigt einen HTML-Syntaxbaum eines Teils einer Suchergebnisliste.

[0027] Im folgenden werden die bevorzugten Ausführungsbeispiele im Detail beschrieben. Die allgemeine Funktion der bevorzugten Ausführungsbeispiele ist in **Fig.** 1 dargestellt. Bevor jedoch mit der Beschreibung weiter fortgefahren wird, werden mehrere Punkte der bevorzugten Ausführungsbeispiele diskutiert.

[0028] In den bevorzugten Ausführungsbeispielen bezieht sich "Primärsuchmaschine" auf eine Internetsuchmaschine, die Informationen aus einer speziellen Datenbank von Internetdokumenten herausholt. Im Gegensatz dazu bezieht sich der Begriff "Meta-Suchmaschine" auf eine Suchmaschine, die keinen direkten Zugang zu solch einer Datenbank besitzt, sondern vielmehr als Schnittstelle zu anderen Primärsuchmaschinen dient. Deshalb umfaßt eine Meta-Suchmaschine eine Schnittstelle zum Benutzer und eine Schnittstelle zu anderen Primärsuchmaschinen, wobei letztere entweder ein Teil der Meta-Suchmaschine ist oder eine getrennte Softwarekomponente ist, die an anderer Stelle im Netzwerk lokalisiert ist.

[0029] Der Begriff "Suchantwortdarstellung" bezieht sich auf das allgemeine Layout des Dokuments, welches das Suchergebnis einer Primärsuchmaschine enthält, jedoch nicht auf eine spezielle Suchantwort, die sich auf eine spezielle Suchanfrage bezieht. Die Darstellung von Suchantworten von Primärsuchmaschinen ist Änderungen unterworfen. Deshalb bezieht sich der Begriff "neue Suchantwortdarstellung" nicht nur auf Suchantwortdarstellungen neuer Primärsuchmaschinen, die zur Meta-Suchmaschine hinzugefügt werden, sondern auch auf Änderungen der Suchantwortdarstellungen von Primärsuchmaschinen, die schon Teil der Meta-Suchmaschine sind.

[0030] Der Begriff "Treffer" bezieht sich auf ein spezielles Dokument, das von der Primärsuchmaschine während der

Internetsuche gefunden wurde. In der Regel sind die von einer Primärsuchmaschine gefundenen Treffer in der Suchantwortdarstellung zwischen anderen Informationen eingebettet. Die mit einem Treffer assoziierten Suchergebnisinformationen sind in einem "Ergebnisrahmen" gruppiert. Da eine Primärsuchmaschine gewöhnlich während einer Internetsuche mehrere Treffer findet, umfaßt die Suchantwortdarstellung mehrere Ergebnisrahmen mit den entsprechenden Treffern und zusätzliche Teile, die sich nicht auf eine spezielle Suchanfrage beziehen.

[0031] In den bevorzugten Ausführungsbeispielen sind die Suchantworten in einer der beiden Markup-Sprachen HTML oder XML codiert. In diesen Sprachen codierte Dokumente können als Sequenz von Markups (Tags) betrachtet werden, die im Text plaziert werden und das Format und Layout des Textes definieren. In diesem Zusammenhang bezieht sich der Begriff "Syntax" und entsprechend "Syntaxelement" auf die Darstellung dieser Markups im Text und ihre spezielle Bedeutung. Ein Syntaxmuster ist eine bestimmte Sequenz solcher Syntaxelemente, wobei die Reihenfolge und das Verhältnis zwischen den Syntaxelementen wichtige Merkmale des Musters sind. Das Ändern eines Syntaxelementes innerhalb eines HTML oder XML Dokuments hat in der Regel Auswirkungen auf die Darstellung des entsprechenden Textabschnitts bezüglich dessen Layout oder Format, wenn es mit einem HTML oder XML kompatiblen Browser (Software-Werkzeug zum Anzeigen von in HTML oder XML codierten Internetdokumenten) angezeigt wird.

[0032] Der Begriff "Suchergebnisinformation" faßt in diesem Zusammenhang die Informationen zusammen, die mit einem von einer Primärsuchmaschine gefundenen Treffer assoziiert sind, insbesondere die URL, den Titel des Dokuments, eine kurze Beschreibung des Inhalts des Dokuments, ein Datum, usw.

[0033] Ein Aspekt des offenbarten Verfahrens zum automatischen Anpassen einer Schnittstelle zwischen einer Meta-Suchmaschine und Primärsuchmaschinen an eine neue Suchantwortdarstellung ist das Erkennen von sich wiederholenden Syntaxmustern in HTML oder XML Dokumenten. Dieser spezielle Aspekt ist allgemein auf die automatische Analyse, die Informationsgewinnung und die Detektion von Formatänderungen in Dokumenten anwendbar. Eine Anwendung zum Beispiel, in der die Erkennung von sich wiederholenden Mustern in der Syntaxstruktur von HTML oder XML Dokumenten vorteilhaft ist, ist das Ausfindigmachen von Preisinformationen in Produktlisten bei E-Business Anwendungen. Deshalb behalten wir uns hiermit die Rechte vor, Schutz für diesem Aspekt ohne Bezug zu Suchmaschinen getrennt zu beanspruchen.

[0034] Obwohl das offenbarte Verfahren vorzugsweise mittels Software implementiert wird, könnte es ebenso ganz oder in Teilen mittels Firmware oder Hardware realisiert werden, ohne daß dabei vom Umfang oder der Idee der Erfindung abgewichen wird.

[0035] Die automatische Anpassung an neue Suchantwortdarstellungen kann auf zwei verschiedene Arten gesehen werden. Von einem Standpunkt aus paßt sich die Meta-Suchmaschine als Ganzes an, um mit neuen Suchantwortdarstellungen umgehen zu können. Von einem anderen, spezielleren Standpunkt aus, paßt nur der Teil der Meta-Suchmaschine, der als Schnittstelle zu anderen Primärsuchmaschinen dient, seine Konfiguration automatisch an neue Suchantwortdarstellungen an. Zu letzterem sind verschiedene Ausführungsbeispiele möglich. In den bevorzugten Ausführungsbeispielen ist die gesamte Schnittstelle in die Meta-Suchmaschine integriert, während in anderen Ausführungsbeispielen die gesamte Schnittstelle oder Teile der Schnittstelle getrennt von der Meta-Suchmaschine und im Netzwerk verteilt sind. Zum Beispiel ist es möglich, nur den Teil der Schnittstelle zu delokalisieren, der die Erkennung und Analyse der Ergebnisrahmen in neuen Suchantwortdarstellungen durchführt.

[0036] Neben der Funktion, sich an neue Suchantwortdarstellungen anzupassen, hat die Schnittstelle die Funktion, die Suchergebnisinformationen aus "alten Suchantworten", d. h. Suchantworten die der Schnittstelle schon bekannt sind, zu extrahieren. Deshalb ist in den bevorzugten Ausführungsbeispielen der erste Schritt, zu bestimmen, ob die fragliche Suchantwort "alt" oder "neu" ist, d. h. ob die Suchergebnisinformationen direkt extrahiert werden können, indem eine, der Schnittstelle schon bekannte Suchantwortdarstellung, verwendet wird, oder ob ein Verfahren gestartet werden muß, um die Suchergebnisinformationen innerhalb der neuen Suchantwortdarstellung zu detektieren. Im allgemeinen macht es die Anpassung an neue Suchantwortdarstellungen jedoch nicht erforderlich, zwischen neuen und "alten" Suchantwortdarstellungen zu unterscheiden, denn es ist durchaus möglich, alle Suchantworten als neu zu betrachten und die Anpassung anzuwenden. Daher wird in anderen Ausführungsbeispielen (nicht gezeigt) das Verfahren zur Detektion der Suchergebnisinformationen innerhalb der Suchantwortdarstellung auf alle Suchantworten angewandt, oder mit anderen Worten, es ist kein Verfahrensschritt implementiert, um Suchantwortdarstellungen automatisch als neu zu erkennen.

[0037] In den bevorzugten Ausführungsbeispielen werden schon erkannte Suchantwortdarstellungen, in der Schnittstelle gespeichert und sind daher der Schnittstelle bekannt. Deshalb ist das Kriterium, eine Suchantwortdarstellung als neu anzusehen, daß sie vorher noch nicht erkannt wurde, d. h. daß sie noch nicht in der Schnittstelle gespeichert ist.

[0038] Normalerweise beinhalten Teile der Suchantwort Informationen, die keinen Bezug zu den Suchergebnisinformationen haben und daher für die weitere Erkennungsanalyse eliminiert werden. Deshalb umfassen die bevorzugten Ausführungsbeispiele einen Verfahrensschritt, der diejenigen Teile der Suchantwort, die die Suchergebnisinformationen tragen, d. h. die Ergebnisrahmen, von denjenigen Teilen, die sich nicht auf ein spezielles Suchergebnis beziehen, automatisch unterscheidet. Das wird höchst bevorzugt dadurch erreicht, daß entweder einer der beiden oder beide Teile automatisch detektiert werden, indem zwei Suchantworten von unterschiedlichen Anfragen verglichen werden. Normalerweise sind die Teile der Suchantwort, die keine Suchergebnisinformationen enthalten, wie Logos, Werbung, Benutzerhinweise oder Kontrollelemente, in den Suchantworten zweier aufeinanderfolgender Suchanfragen identisch und werden bevorzugter Weise als diejenigen Teile identifiziert, deren Inhalt sich nicht in zwei unterschiedliche Suchantworten ändert.

[0039] In der Regel findet die Primärsuchmaschine auf eine einzige Suchanfrage mehrere Treffer. Diese Treffer werden als Liste von Ergebnisrahmen, die jeweils einen Treffer enthalten, angezeigt. Eine Möglichkeit, Ergebnisrahmen innerhalb der Suchantwortdarstellung zu erkennen, ist, dieses wiederholte Auftreten der Ergebnisrahmen zu nutzen.

[0040] Deshalb verwendet das bevorzugte Ausführungsbeispiel zur Detektion der Ergebnisrahmen Suchantworten, die mehr als einen Ergebnisrahmen enthalten. Diese sich wiederholenden Ergebnisrahmen werden höchst bevorzugt aufgrund ihres ähnlichen Aussehens innerhalb der Suchantwortdarstellung identifiziert.

[0041] Normalerweise umfassen die in einem Ergebnisrahmen enthaltene Suchergebnisinformationen mehrere Komponenten, wie die URL, den Titel, eine kurze Beschrei-

bung und das Datum des entsprechenden Dokuments. Diese verschiedenen Komponenten werden in verschiedenen Formaten und Layouts angezeigt und werden daher höchst bevorzugt aufgrund ihres speziellen visuellen Aussehens identifiziert.

[0042] Das Blockdiagramm der **Fig. 1** zeigt für die bevorzugten Ausführungsbeispiele der Erfindung die Funktion einer Meta-Suchmaschine **4** als Schnittstelle zwischen einem Benutzer-Hostcomputer **2** und mehreren Servern von Primärsuchmaschinen **6**. Anstatt eine separate Suchanfrage zu allen Servern der Primärsuchmaschinen **6** zu schicken, richtet der Benutzer-Hostcomputer **2** seine Anfrage nur einmal an den Meta-Suchmaschinenserver **4**, der die Anfrage an die speziellen Anforderungen der Primärsuchmaschinen **6** anpaßt und die speziellen Suchanfragen an die einzelnen Server der Primärsuchmaschinen **6** übermittelt. Nachdem die Meta-Suchmaschine die individuellen Suchergebnisse von den Primärsuchmaschinen empfangen hat, detektiert und bündelt sie die Suchergebnisse, konvertiert sie in ein einheitliches Format und schickt sie zurück zum Benutzerhost. Dadurch ist der Benutzer in der Lage, durch Senden nur einer einzigen Suchanfrage auf mehrere Primärsuchmaschinen gleichzeitig zuzugreifen und die gefilterten und vereinheitlichten Suchantworten der verschiedenen Primärsuchmaschinen alle auf einmal am Bildschirm zu erhalten.

[0043] **Fig. 2** ist eine funktionelle Darstellung auf hoher Ebene einer Schnittstelle **8** zwischen der Meta-Suchmaschine **4** und der Primärsuchmaschine **6**. Die Schnittstelle **8** dient im allgemeinen als Konfigurationseinheit, um die Meta-Suchmaschine **4** an neue Suchantwortdarstellungen anzupassen. Wie oben bereits erwähnt, kann die Schnittstelle **8** entweder als Teil der Meta-Suchmaschine **4** oder als getrennte Softwarekomponente implementiert werden. Das Schema der **Fig. 2** nimmt an, daß von der Meta-Suchmaschine **4** eine Suchanfrage eines Benutzers an die Primärsuchmaschine **6** weitergeleitet wurde und daß die Primärsuchmaschinen **6** ihre individuellen Suchantworten gefunden haben. Die Suchantworten sind beispielsweise in der HTML Markup-Sprache codiert.

[0044] Als Antwort auf die Suchanfrage der Meta-Suchmaschine **4** schickt die Primärsuchmaschine **6** die Suchantwort via HTTP in Form eines HTML Dokuments zum Interface **8** zurück. Eine Suchergebniserkennung **12**, welche die Suchergebnisrahmen und ihren Inhalt in den Suchantworten detektiert, wird auf die individuellen HTML Suchantworten der Primärsuchmaschinen angewandt. Die Suchergebniserkennung **12** ist in der Lage, die Ergebnisrahmen zu detektieren und die mit dem Treffer assoziierten Informationen zu extrahieren, auch wenn das Layout, das Format oder die Position der Ergebnisrahmen innerhalb des HTML Dokuments oder die interne Struktur des Ergebnisrahmens verändert sind oder eine komplett neue Primärsuchmaschine, die eine neue Suchantwortdarstellung verwendet, zu den existierenden Primärsuchmaschinen hinzugefügt wird. Die in einem Ergebnisrahmen enthaltenen Ergebnisinformationen **14**, umfassen die URL, den Titel des referenzierten Dokuments, eine kurze Beschreibung des Inhalts des referenzierten Dokuments, das Datum, die Quelle des Sucherergebnisses, d. h. der Name der Primärsuchmaschine, und eine Wertung, die die Relevanz des gefundenen Dokuments angibt. Diese extrahierten Suchergebnisinformationen **14** werden dann weiter zur Meta-Suchmaschine **4** übertragen. In den bevorzugten Ausführungsbeispielen werden die Treffer gemäß ihrer Wertung klassifiziert und in einem einheitlichen Format angezeigt, wobei Treffer die von mehr als einer Primärsuchmaschine gefunden wurden, entfernt werden.

[0045] In den bevorzugten Ausführungsbeispielen können zwei Fälle der Extraktion von Suchergebnissen unterschieden werden. Erstens, die fragliche Suchantwortdarstellung ist bereits bekannt und in die Schnittstelle integriert, und zweitens, die Suchantwortdarstellung ist neu. Da die Meta-Suchmaschine keine Kontrolle über das Layout der Primärsuchmaschinen hat und nicht einmal von der Primärsuchmaschine über solche Layout- oder Formatänderungen benachrichtigt wird, muß sie in der Lage sein, beide Fälle, also "alte" und "neue" Suchantwortdarstellungen zu handhaben. Die Extraktion **18** von Suchergebnisinformationen aus neuen Suchantwortdarstellungen und wie eine Suchantwortdarstellung als neu erkannt wird, wird in **Fig. 4** für die bevorzugten Ausführungsbeispiele ausführlicher erläutert. In **Fig. 3** werden die Verfahrensschritte **16** zur Extraktion der Suchergebnisinformationen aus Suchantworten, die "alte", der Schnittstelle bereits bekannte Suchantwortdarstellungen haben, veranschaulicht.

[0046] In den in **Fig. 3** dargestellten bevorzugten Ausführungsbeispielen wird ein HTML Suchantwortdokument **10** von einer Primärsuchmaschine **6** zurückgegeben. Ein hierarchischer HTML Syntaxbaum, der als Basis für alle weiteren Verarbeitungsschritte dient, wird von einem Syntaxbaum-Generator **20** erstellt.

[0047] Ein Extraktionsschritt **22**, der zwei Verfahrensschritte umfaßt, nämlich die Extraktion **21** des Ergebnisrahmens und die Extraktion der Suchergebnisinformationen **23**, wird auf den Syntaxbaum der Suchantwort angewandt. Zuerst lokalisiert und extrahiert die Ergebnisrahmenextraktion unterschiedliche, in der Suchantwort enthaltene Ergebnisrahmen **25**, indem sie den Syntaxbaum der Suchantwort mit bekannten Syntaxmustern von in einer Datenbank **40** gespeicherten Ergebnisrahmen vergleicht. Dazu wird ein Syntaxmuster der Datenbank **40** mit allen Syntaxteilbäumen der Suchantwort verglichen. Wenn ein Syntaxteilbaum mit dem Syntaxmuster identisch ist, wird ein Ergebnisrahmen detektiert. Wenn alle Syntaxteilbäume der Suchantwort getestet wurden, werden die detektierten Ergebnisrahmen **25** zum zweiten Extraktionsschritt **23** übermittelt, um die Suchergebnisinformationen zu identifizieren. Zusammen mit dem Syntaxmuster des Ergebnisrahmens wird in den bevorzugten Ausführungsbeispielen auch die Rolle der Syntaxelemente als Träger der Suchergebnisinformationen in der Datenbank **40** als ein zusätzliches Attribut gespeichert. Dadurch wird ein bestimmter Typ der Suchergebnisinformationen **42** (URL, Titel, Datum, Beschreibung, Quelle, Wertung) mit einem speziellen Syntaxelement des Ergebnisrahmens assoziiert.

[0048] Diese Attribute werden dann von dem Extraktionsschritt **23** verwendet, um die Suchergebnisinformationen **42** des Treffers zu bestimmen und der Meta-Suchmaschine **4** weiterzuleiten. In anderen Ausführungsbeispielen (nicht gezeigt) wird die Verknüpfung der Suchinformationen mit speziellen Syntaxelementen nicht als Attribute in der Datenbank zusammen mit dem Syntaxmuster gespeichert, sondern werden in jedem Extraktionsschritt **22** identifiziert.

[0049] In **Fig. 4** sind die Schritte der von der Schnittstelle **8** durchgeführten automatischen Erkennung der Suchergebnisinformationen in einem Blockdiagramm dargestellt. Die komplette Schnittstelle umfaßt zwei Extraktionsteile **16** und **18**, wobei die Extraktion **16** Suchantworten von "alten" Suchantwortdarstellungen verarbeitet, die zuvor erkannt worden sind und schon in die Schnittstelle integriert wurden. Auf der anderen Seite führt die Extraktion **18** eine Erkennung von neuen Suchantwortdarstellungen durch, die der Schnittstelle noch nicht bekannt sind.

[0050] Wenn man als Eingabe der Schnittstelle **8** ein HTML Suchergebnisdokument **10** annimmt, das von einer Primärsuchmaschine **6** zurückgegeben wurde, analysiert der Syntaxbaum-Generator **20** die HTML Syntaxstruktur des

Suchergebnisdokuments, indem er die HTML Tags innerhalb des Dokuments erkennt und einen hierarchischen HTML Syntaxbaum erstellt, der das hierarchische Verhältnis der Syntaxelemente (Tags) repräsentiert. Das HTML Dokument wird so in einen Syntaxbaum transformiert, der das Format und die Layoutstruktur der ursprünglichen HTML Suchantwort repräsentiert.

[0051] Das Ziel des Extraktionsschrittes **23** ist festzustellen, ob der fragliche HTML Syntaxbaum eine Darstellung von Ergebnisrahmen enthält, die dem System schon bekannt sind. Um dies zu erreichen, wird der HTML Syntaxbaum mit dem HTML Syntaxmuster einer Datenbank **40** verglichen, in dem die HTML Syntaxstruktur des bekannten Ergebnisrahmens gespeichert ist. Wenn der Extraktionsschritt **22** in der Lage ist, die Ergebnisrahmen innerhalb des HTML Suchergebnisdokuments zu lokalisieren, wird die Bedeutung der verschiedenen Syntaxelemente im Ergebnisrahmen bestimmt und die entsprechenden Suchergebnisinformationen **42** extrahiert und zur Meta-Suchmaschine **4** übermittelt. Andernfalls gibt es zwei Möglichkeiten. Erstens, die Primärsuchmaschine hat keine Suchergebnisse gefunden, oder zweitens, die Suchantwortdarstellung ist für die Schnittstelle neu und es konnten aus diesem Grunde keine Ergebnisrahmen extrahiert werden. Um sich für eine der beiden Möglichkeiten zu entscheiden, werden zwei Kriterien überprüft. Erstens, es wird geprüft, ob die Anzahl der aufeinanderfolgenden Fehlversuche, Ergebnisrahmen zu extrahieren, einen gewissen Grenzwert überschreiten, und zweitens, ob Ergebnisrahmen von Testanfragen extrahiert werden können, von denen bekannt ist, daß sie Suchergebnisse finden.

[0052] Wenn im Extraktionsschritt **22** keine Rahmen detektiert werden können, wird ein Zähler **24** um eine Einheit erhöht. Wenn der Zähler unter einem bestimmten Schwellwert **26** liegt, dann wird vermutet, daß die ursprüngliche Suchanfrage keine Suchergebnisse gefunden hat und daher wird eine "kein Suchergebnis" Mitteilung **44** an die Meta-Suchmaschine **4** übermittelt. Andernfalls ist die Ergebnisrahmenextraktion für eine bestimmte Anzahl mißlungen, so daß es daher sehr wahrscheinlich ist, daß die Suchantwortdarstellung neu ist. Daher wird das zweite Kriterium geprüft und es werden einige Testanfragen vom Verfahrensschritt **28** durchgeführt, für die bekannt ist, daß die Primärsuchmaschine mehr als einen Treffer findet. Ein Vergleichsverfahren ähnlich dem im Extraktionsschritt **22** wird im Verfahrensschritt **28** auf die Suchantworten der Testanfragen angewendet. Wenn Ergebnisrahmen von den Suchantworten der Testanfragen extrahiert werden können, was bedeutet, daß im Gegensatz zur ersten Annahme gemäß dem ersten Kriterium, die Suchantwortdarstellung nicht neu ist und Ergebnisrahmen generell von dieser Suchantwortdarstellung extrahiert werden können, dann wird vermutet, daß die ursprüngliche Suchanfrage keine Suchergebnisse gefunden hat. Daher wird die "kein Ergebnis" Mitteilung **44** zur Meta-Suchmaschine **4** übermittelt. Wenn jedoch die Extraktion **28** keine Ergebnisrahmen aus den Suchantworten der Testanfragen extrahieren konnte, wird schließlich davon ausgegangen, daß die Suchantwortdarstellung neu ist und der Teil **18** der Schnittstelle wird durch den Verfahrensschritt **30** initialisiert, um die neue Suchantwortdarstellung zu erkennen. Insgesamt geht die Schnittstelle von einer neuen Suchantwortdarstellung aus, wenn beide der folgenden Bedingungen zutreffen: 1) die Suchrahmenextraktion mißlang für eine Reihe von aufeinanderfolgenden Suchanfragen, und 2) die Suchrahmenextraktion mißlang für eine Reihe von Testanfragen. In anderen Ausführungsbeispielen (nicht gezeigt) wird nur die erste Bedingung verwendet, um die Erkennung von neuen Suchantwortdarstellungen zu initialisieren.

[0053] Zur Erkennung von neuen Suchantwortdarstellungen fordert Verfahrensschritt **32** zwei verschiedene Testanfragen von der Primärsuchmaschine **6** an, von denen bekannt ist, daß sie für jede der Testanfragen mehrere Treffer ergeben. Der Verfahrensschritt **34** vergleicht dann die Syntaxbäume der Suchantworten der beiden Testanfragen, und identifiziert diejenigen Teile des Syntaxbaumes (Teilbäume), die in beiden Syntaxbäumen vollkommen identisch sind. Da man davon ausgeht, daß diese Teilbäume keine Suchergebnisinformationen enthalten, wie Werbung oder Kontrollelemente, werden sie vom HTML Syntaxbaum der Suchantworten der Testanfragen entfernt. Experimentellen Daten zufolge kann die Größe des HTML Syntaxbaumes durch diese Hintergrundbeseitigung **34** um etwa 40% reduziert werden.

[0054] Der reduzierte HTML Syntaxbaum wird zum Verfahrensschritt **36** weitergeleitet, der eine Clusteranalyse durchführt, um innerhalb der Suchantwortdarstellung die HTML Syntaxstruktur des Ergebnisrahmens zu erkennen. Die Clusteranalyse **36** detektiert in der Syntaxbaumstruktur der Suchantwortdarstellung sich wiederholende Muster und identifiziert diese als die Syntaxstruktur der Ergebnisrahmen. Das Ergebnis Clusteranalyse ist also ein HTML Syntaxmuster, das den Ergebnisrahmen darstellt. Im Verfahrensschritt **38** werden die Ergebnisinformationen den verschiedenen Syntaxelementen des Ergebnisrahmens zugeordnet. Die Bestimmung der Bedeutung eines bestimmten Syntaxelements wird typischer Weise durch Anwendung heuristischer Kriterien durchgeführt: 1) die URL wird durch ein spezielles HTML Tag <a href = "http/ /. . ."> erkannt, 2) die Beschreibung durch den längsten einheitlichen Textbereich, 3) der Titel durch das den Fettdruck definierende Tag <b> und ein umgebendes <a href = ". . ."> Tag, 4) das Datum durch ein Zahlenformat, und 5) die Wertung durch die Textmarke "%" und die Reihenfolge der Treffer innerhalb der Suchantwort. Ein anderes Kriterium, das berücksichtigt wird, ist die Reihenfolge der Elemente innerhalb des Ergebnisrahmens.

[0055] Schließlich wird das extrahierte HTML Syntaxmuster, das die Zuordnung der Syntaxelemente zu den verschiedenen Bestandteilen der Suchergebnisinformationen enthält, der Datenbank **40**, die bereits erkannte HTML Syntaxmuster von Ergebnisrahmen enthält, hinzugefügt.

[0056] Dasselbe, oben beschriebene Verfahren wird der Reihe nach mit den Suchantworten der anderen Primärsuchmaschinen **6** und ihren HTML Suchergebnisdokumenten durchgeführt.

[0057] Die Clusteranalyse **36** der bevorzugten Ausführungsbeispiele wird unten im einzelnen beschrieben. In **Fig. 5** ist eine typische, in einem Internetbrowser angezeigte Suchantwort der bekannten Altavista-Primärsuchmaschine, gezeigt. Die Suchergebnisseite **46** zeigt Teile der Suchantwort, die mit der Suchanfrage in Beziehung stehende Informationen enthalten, nämlich die Suchergebnisliste **48**, und andere Teile (**50**, **54**), die nicht in Bezug zu einer speziellen Suchanfrage stehen. Letztere umfassen Werbung **50**, Kontrollelemente **52**, Logos **54** und Benutzerhinweise **56**. Andererseits umfaßt die Ergebnisliste **48** eine aufeinanderfolgende Anordnung von Ergebnisrahmen **58**, welche die URL des entsprechenden Treffers **60**, den Titel **62**, eine kurze Beschreibung des Inhalts des referenzierten Dokuments **64** und das Datum **66** enthalten.

[0058] **Fig. 6** zeigt einen zu einem speziellen Suchergebnisrahmen **58** gehörenden Ausschnitt eines HTML Quellcodes **68**. Dieser Ausschnitt **68** setzt sich aus HTML Syntaxelementen (Tags) zusammen, die das Format und das Layout des enthaltenen Textes definieren und den Text **72** des Ausschnitts des Dokuments selbst. Zum Beispiel definiert das Syntaxelement <dl> einen bestimmten Listentyp <dl> defi-

niert ein Element dieser Liste, <dd> definiert den Inhalt des Listenelements, <b> bewirkt, daß der nachfolgende Text fett gedruckt ist,
 fügt einen Zeilenumbruch ein und <a href = ". . ."> ist ein Querverweis auf eine URL, wobei jedes der Elemente sein entsprechendes Endtag **74** </dl>, </dt>, usw. aufweist.

[0059] **Fig.** 7 zeigt den von dem Syntaxbaum-Generator **20** erzeugten HTML Syntaxbaum, der zum Suchergebnisrahmen **68** der **Fig.** 6 gehört. Der Syntaxbaum bildet die Basis für alle weiteren Verfahrensschritte. In den bevorzugten Ausführungsbeispielen wird der Syntaxbaum-Generator von einem Modul der Interpreterprogrammiersprache PERL ausgeführt. Die HTML Tags werden in dem hierarchischen Syntaxbaum **76** in der Reihenfolge ihres Auftretens in den Dokumenten angeordnet, während ihre Abhängigkeit von anderen Tags **70** durch ihre Ebene **78** wiedergegeben wird. Wenn ein bestimmtes Tag angewendet wird, bevor das Endtag des vorangegangenen Tags gesetzt ist, dann wird das betreffende Tag im Syntaxbaum eine Ebene tiefer klassifiziert. Der zu einem bestimmten Suchergebnisrahmen gehörende Syntaxteilbaum **76** zum Beispiel beginnt auf der Ebene **7** und geht runter bis Ebene **9**.

[0060] Schließlich wird jeder Knoten des HTML Syntaxbaumes durch die folgenden drei Attribute charakterisiert, den Typ des Tags **70**, die passende Ebene **78** und seine aufeinanderfolgende Position innerhalb des HTML Dokuments. Diese drei Attribute spannen den in **Fig.** 8 dargestellten Merkmalsraum **80** auf. Jedes Syntaxelement wird in dem dreidimensionalen Merkmalsraum entsprechend den drei Dimensionen Typ des Tags **82**, Ebene **84** innerhalb der hierarchischen Syntaxstruktur und der Position **86** innerhalb des HTML Dokuments klassifiziert. In **Fig.** 8 ist eine Anordnung mehrerer aufeinanderfolgender Syntaxelemente **76** dargestellt, die im Merkmalsraum ein Muster bilden und ein Teil eines speziellen Ergebnisrahmens sind.

[0061] Der Syntaxbaum von Teilen einer Suchergebnisliste ist in **Fig.** 9 gezeigt. In den bevorzugten Ausführungsbeispielen besteht das Verfahren zur Detektion des Suchergebnisrahmens darin, innerhalb des Syntaxbaumes **90** nach gleichen Clustern (Mustern) von Syntaxelementen der Größe **5** zu suchen. Die Lokalisierung dieser Cluster **88** erlaubt es dann, die verschiedenen Suchergebnisrahmen voneinander zu unterscheiden und die Syntaxstruktur eines solchen Ergebnisrahmens zu bestimmen. Dieses Syntaxmuster wird dann in der Datenbank **40** der **Fig.** 3 gespeichert, wo es dazu verwendet wird, Suchergebnisrahmen aus zukünftigen Suchantworten zu extrahieren.

[0062] In **Fig.** 7 ist das Syntaxmuster eines Ergebnisrahmens mit seinen absoluten Ebenen gezeigt. In anderen Ausführungsbeispielen (nicht gezeigt) werden die detektierten Syntaxmuster einer neuen Suchantwortdarstellung normalisiert in der Datenbank **40** der **Fig.** 3 gespeichert, d. h. die oberste Ebene des hierarchischen Teilbaums, der dem detektierten Syntaxmuster entspricht, wird auf 1 gesetzt und die Ebenen der nachfolgenden Syntaxelemente werden entsprechend angepaßt. Folglich wird nur die Ebene der Elemente des Syntaxmusters relativ zur obersten Ebene gespeichert, was das Vergleichsverfahren **22** invariant gegenüber der absoluten Ebene des Syntaxmusters des Ergebnisrahmens innerhalb der Suchantwortdarstellung macht.

[0063] Eine Bedingung, die an die Detektion der Ergebnisrahmen gestellt wird, um das Verfahren **18** verläßlicher zu machen, ist, daß von der Syntaxstruktur eines Ergebnisrahmens eine gewisse Komplexität gefordert wird, das heißt, eine minimale Anzahl Tags und eine minimale Tiefe der Ebenen (tiefste Ebene des normalisierten Syntaxmusters).

[0064] In einigen Primärsuchmaschinen ist die Syntaxstruktur der Ergebnisrahmen nicht für alle Ergebnisrahmen identisch, sondern variiert innerhalb einer Suchantwortdarstellung leicht. Deshalb ist ein Modell erforderlich, das ähnliche Ergebnisrahmen in nur einem Muster repräsentiert. Zum Beispiel kann ein zusätzliches Syntaxelement, das einen Zeilenumbruch definiert, in das Syntaxmuster eingefügt werden. Dafür wird ein Platzhalter, der den Typ des Tags unbestimmt läßt, an der Stelle des Syntaxmusters einfügt, wo möglicherweise ein zusätzlicher Tag vorkommt. Während des Syntaxmustervergleichs des Extraktionsschritts **22** der **Fig.** 3 kann das dem Platzhalter entsprechende Syntaxbaumelement jeden Syntaxtyp annehmen. Das erweiterte Syntaxmuster ist daher so flexibel, daß es ähnliche aber nicht identische Ergebnisrahmen in einer Suchantwortdarstellung detektiert.

[0065] Es ist daher ein allgemeiner Zweck der offenbarten Ausführungsbeispiele, ein verbessertes Verfahren, Computersystem und Computerprogramm-Produkt zur Verfügung zu stellen, um ein Interface einer Meta-Suchmaschine automatisch, d. h. ohne manuellen Eingriff, an eine neue Suchantwortdarstellung anzupassen, wohingegen im Stand der Technik die Kontrolle und Anpassung manuell vorgenommen wird.

[0066] Alle Veröffentlichungen und existierenden Systeme, die in dieser Beschreibung erwähnt werden, sind per Bezug hier miteinbezogen.

[0067] Auch wenn bestimmte Verfahren, Systeme und Produkte, die gemäß der Lehre der Erfindung erstellt sind, hier beschrieben wurden, beschränkt sich der Bereich dieses Patents nicht darauf. Im Gegenteil, dieses Patent schließt alle Ausführungsbeispiele der Lehre der Erfindung mit ein, die entweder wörtlich oder unter der Doktrin der Äquivalenz in den Bereich der beigefügten Ansprüche fallen.

Legende zu den Figuren

Englisch	Deutsch
Fig. 1:	
user host	Benutzerhost
Meta Search Engine (MSE)	Meta-Suchmaschine (MSM)
Servers of Pimary Search Engines (PSE)	Server der Primärsuchmaschinen (PSM)
Fig. 2:	
Primary Search Engines	Primärsuchmaschinen
Search result recognition	Suchergebniserkennung
URL (Uniform Resource Locator)	URL (Querverweis auf Web-Site)
Title	Titel
Date	Datum
Description	Beschreibung
Interface	Schnittstelle
Meta search engine	Meta-Suchmaschine
Fig. 3:	
Start PSE	Start PSM
HTML search result	HTML Suchergebnis
Syntax tree parser	Syntaxbaum-Generator
Data base	Datenbank

Fig. 3:

result frame extraction	Ergebnisrahmenextraktion
result frames	Ergebnisrahmen
extraction of search result information	Extraktion der Suchergebnisinformation
URL (Uniform Resource Locator)	URL (Querverweis auf Web-Site)
Title	Titel
Description	Beschreibung
Date	Datum
MSE	MSM

Fig. 4:

Start PSE	Start PSM
HTML search result	HTML Suchergebnis
Syntax tree parser	Syntaxbaum-Generator
Database of HTML Syntax patterns	Datenbank der Syntaxmuster
Extraction successful	Extraktion erfolgreich
URL (Uniform Resource Locator)	URL (Querverweis auf Web-Site)
Title	Titel
Description	Beschreibung
Date	Datum
Score	Wertung
no result frames	keine Ergebnisrahmen
Assignment of result information type	Zuordnung des Ergebnisinformationstyps
Cluster Analyse	Clusteranalyse
number of search queries with no frames	Anzahl der Suchanfragen ohne Rahmen
Background elimination	Hintergrundbeseitigung
false	falsch
true	wahr
2 different test queries	2 verschiedene Testanfragen
no search results	keine Suchergebnisse
Initialize Recognition of new representation failed	Initialisiere Erkennung einer neuen Darstellung mißlungen
Extraction of test queries	Extraktion der Testanfragen
MSE display result list	MSM Anzeige Ergebnisliste

Fig. 5:

search string	Suchtext
Go	Start
Title	Titel
Description	Beschreibung
Date	Datum
Commercial	Werbung

Fig. 8:

type of tag	Typ des Tag
level	Ebene
position	Position

Patentansprüche

1. Verfahren, welches mit einer Meta-Suchmaschine durchgeführt wird, bei dem eine von einer Primärsuchmaschine in einer Suchantwortdarstellung geliefert Suchantwort von der Meta-Suchmaschine verarbeitet wird, wobei das Verfahren umfaßt: die Meta-Suchmaschine paßt sich selbst an eine neue Suchantwortdarstellung an.

2. Verfahren nach Anspruch 1, bei dem die Meta-Suchmaschine eine Schnittstelle zum Extrahieren von Suchergebnissen aus der Suchantwort umfaßt und die Anpassung der Meta-Suchmaschine durch automatisches Konfigurieren der Schnittstelle für die neue Suchantwortdarstellung durchgeführt wird.

3. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die Meta-Suchmaschine automatisch eine neue Suchantwortdarstellung erkennt.

4. Verfahren nach einem der vorhergehenden Ansprüche, bei dem eine Suchantwortdarstellung als neu betrachtet wird, wenn die Meta-Suchmaschine sie zuvor nicht erkannt hat.

5. Verfahren nach einem der vorhergehenden Ansprüche, bei dem das sich Anpassen der Meta-Suchmaschine des weiteren umfaßt, daß mindestens eines der beiden automatisch detektiert wird:

diejenigen Teile einer Suchantwortdarstellung, die keine Suchergebnisinformationen enthalten, und

ii) Ergebnisrahmen in einer Suchantwortdarstellung, wobei Ergebnisrahmen diejenigen Teile einer Suchantwortdarstellung sind, welche die Suchergebnisinformationen enthalten.

6. Verfahren nach einem der vorhergehenden Ansprüche, bei dem das automatische Detektieren derjenigen Teile einer neuen Suchantwortdarstellung, die keine Suchergebnisinformationen enthalten, des weiteren das Vergleichen von mindestens zwei verschiedenen Suchantworten umfaßt.

7. Verfahren nach einem der vorhergehenden Ansprüche, bei dem das automatische Detektieren der genannten Teile des weiteren umfaßt, daß Teile, die keine Informationen enthalten, als diejenigen Teile identifiziert werden, deren Inhalt sich in verschiedenen Suchantworten nicht ändert.

8. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem das Detektieren von Ergebnisrahmen in Suchantworten des weiteren das Analysieren von Suchantworten, die mehr als ein Ergebnisrahmen enthalten, umfaßt, wobei Ergebnisrahmen diejenigen Teile einer Suchantwortdarstellung sind, welche die Suchergebnisinformationen enthalten.

9. Verfahren nach einem der vorhergehenden Ansprüche, bei denen das Detektieren von Ergebnisrahmen in Suchantworten des weiteren umfaßt, daß Teile der Suchantwort identifiziert werden, die ein ähnliches Aussehen haben.

10. Verfahren nach einem der vorhergehenden Ansprüche, welches Komponenten eines Ergebnisrahmens verwendet, wobei das Aussehen der verschiedenen Komponenten eines Ergebnisrahmens dazu verwendet wird, um die spezielle Art der Information, welche die entsprechende Komponente enthält, zu identifizieren, wobei Ergebnisrahmen diejenigen Teile einer Suchantwortdarstellung sind, welche die Suchergebnisinformationen enthalten.

11. Verfahren, welches von einem Computersystem durchgeführt wird, zum Konfigurieren einer Schnittstelle zu mindestens einer Primärsuchmaschine, um

Suchergebnisse aus einer von der Primärsuchmaschine in einer Suchantwortdarstellung gelieferten Suchantwort zu extrahieren, wobei das Verfahren ein automatisches Anpassen der Schnittstelle an eine neue Suchantwortdarstellung umfaßt.

12. Verfahren nach Anspruch 11, bei dem die Schnittstelle ein Teil der Meta-Suchmaschine ist.

13. Verfahren nach Anspruch 11 oder 12, bei dem die automatische Anpassung der Schnittstelle angewendet wird, wenn die Suchantwortdarstellung als neu erkannt wird.

14. Verfahren nach einem der Ansprüche 11 bis 13, bei dem eine Suchantwortdarstellung als neu betrachtet wird, wenn die Schnittstelle sie zuvor nicht erkannt hat.

15. Verfahren nach einem der Ansprüche 11 bis 14, bei dem das automatische Anpassen der Schnittstelle des weiteren das automatische Detektieren von mindestens einem des folgenden umfaßt:

i) diejenigen Teile einer Suchantwortdarstellung, die keine Suchergebnisinformationen enthalten, und

ii) Ergebnisrahmen in einer Suchantwortdarstellung.

16. Verfahren nach einem der Ansprüche 11 bis 15, bei dem das Detektieren derjenigen Teile einer neuen Suchantwortdarstellung, die keine Sucherergebnisinformationen enthalten, des weiteren mindestens das Vergleichen von verschiedenen Suchantworten umfaßt.

17. Verfahren nach einem der Ansprüche 11 bis 16, bei dem eine Suchantwortdarstellung durch eine Syntaxstruktur der Suchantwortdarstellung charakterisiert wird.

18. Verfahren nach einem der Ansprüche 11 bis 17, bei dem das automatische Anpassen der Schnittstelle des weiteren umfaßt, daß Ergebnisrahmen durch Detektieren von Mustern in der Syntaxstruktur der Suchantwortdarstellung identifiziert werden.

19. Verfahren nach einem der Ansprüche 11 bis 18, bei dem das Detektieren von Mustern in der Syntaxstruktur der Suchantwortdarstellung des weiteren das Suchen nach wiederholtem Auftreten von Mustern in der Syntaxstruktur umfaßt.

20. Verfahren nach einem der Ansprüche 11 bis 19, bei dem das Detektieren von Mustern in der Syntaxstruktur der Suchantwortdarstellung des weiteren das Suchen nach sich wiederholenden Mustern in einem Merkmalsraum mit mehr als einer Dimension umfaßt, wobei die Merkmale von der Syntaxstruktur der Suchantwortdarstellung abgeleitet werden.

21. Verfahren nach einem der Ansprüche 11 bis 20, bei dem die Suchantwortdarstellung mit einer Markup-Sprache codiert ist.

22. Verfahren nach einem der Ansprüche 11 bis 21, bei dem die Suchantwortdarstellung mit mindestens einem der beiden HTML und XML codiert ist.

23. Verfahren nach einem der Ansprüche 11 bis 22, bei dem das automatische Anpassen der Schnittstelle des weiteren umfaßt, daß die Bedeutung der Teile des Ergebnisrahmens automatisch bestimmt werden.

24. Verfahren nach einem der Ansprüche 11 bis 22, bei dem das Bestimmen der Bedeutung der Teile des Ergebnisrahmens des weiteren umfaßt, daß die Syntaxelemente des Ergebnisrahmens den zugehörigen Suchergebnisinformationen zugeordnet werden.

25. Computersystem umfassend:
eine Meta-Suchmaschine, die eine Schnittstelle zu mindestens einer Primärsuchmaschine umfaßt;
einen Konfigurator;
wobei der Konfigurator dazu ausgebildet ist, um die Schnittstelle automatisch an eine neue Suchantwortdarstellung der Primärsuchmaschine anzupassen.

26. Computersystem nach Anspruch 25, bei dem der Konfigurator ein Teil der Meta-Suchmaschine ist.

27. Computersystem nach Anspruch 25 oder 26, bei dem die Meta-Suchmaschine und der Konfigurator örtlich getrennt und über ein Netzwerk miteinander verbunden sind.

28. Computersystem nach einem der Ansprüche 25 bis 27, bei dem der Konfigurator dazu ausgebildet ist, um Suchergebnisse, die in Ergebnisrahmen einer Suchantwort mit einer neuen Suchantwortdarstellung enthalten sind, automatisch zu extrahieren.

29. Computersystem nach einem der Ansprüche 25 bis 28, bei dem das Detektieren von Ergebnisrahmen in neuen Suchantwortdarstellungen umfaßt, daß in Suchantworten, die mehr als einen Ergebnisrahmen enthalten, diejenigen Teile der Suchantwort, die ein ähnliches Aussehen haben, identifiziert werden.

30. Computerprogramm-Produkt, umfassend einen Programmcode, zum Durchführen eines Verfahrens, das, wenn auf einem Computersystem ausgeführt, dem Konfigurieren einer Schnittstelle zu mindestens einer Primärsuchmaschine dient, um Suchergebnisse aus einer Suchantwort einer Primärsuchmaschine in einer Suchantwortdarstellung zu extrahieren, wobei das Verfahren ein automatisches Anpassen der Schnittstelle an eine neue Suchantwortdarstellung umfaßt.

31. Computerprogramm-Produkt nach Anspruch 30, bei dem der Programmcode auf einem computerlesbaren Datenträger gespeichert ist oder in Form von Signalen über ein Computernetzwerk übertragen wird.

32. Computerprogramm-Produkt nach Anspruch 30 oder 31, bei dem das Anpassen der Schnittstelle das Detektieren von Ergebnisrahmen in neuen Suchantwortdarstellungen umfaßt, wobei in Suchantworten, die mehr als ein Ergebnisrahmen enthalten, diejenigen Teile der Suchantwort, die ein ähnliches Aussehen haben, identifiziert werden.

33. Computerprogramm-Produkt nach einem der Ansprüche 30 bis 32, bei dem eine Programmkomponente zum Konfigurieren der Schnittstelle automatisch neue Suchantwortdarstellungen erkennt und diese Darstellungen speichert, während eine andere Programmkomponente die gespeicherten Darstellungen verwendet, um bekannte Suchantwortdarstellungen zu verarbeiten.

34. Computerprogramm-Produkt nach einem der Ansprüche 30 bis 33, bei dem die Programmkomponente, die neue Suchantwortdarstellungen erkennt, umfaßt, daß die Bedeutung der Teile des Ergebnisrahmens automatisch bestimmt und die zugehörigen Suchergebnisinformationen zugeordnet werden.

35. Computerprogramm-Produkt mit einem oder mehreren Merkmalen einer der vorhergehenden Ansprüche.

Hierzu 9 Seite(n) Zeichnungen

Nummer: DE 100 56 681 A1
Int. Cl.[7]: G 06 F 15/18
Offenlegungstag: 23. Mai 2002

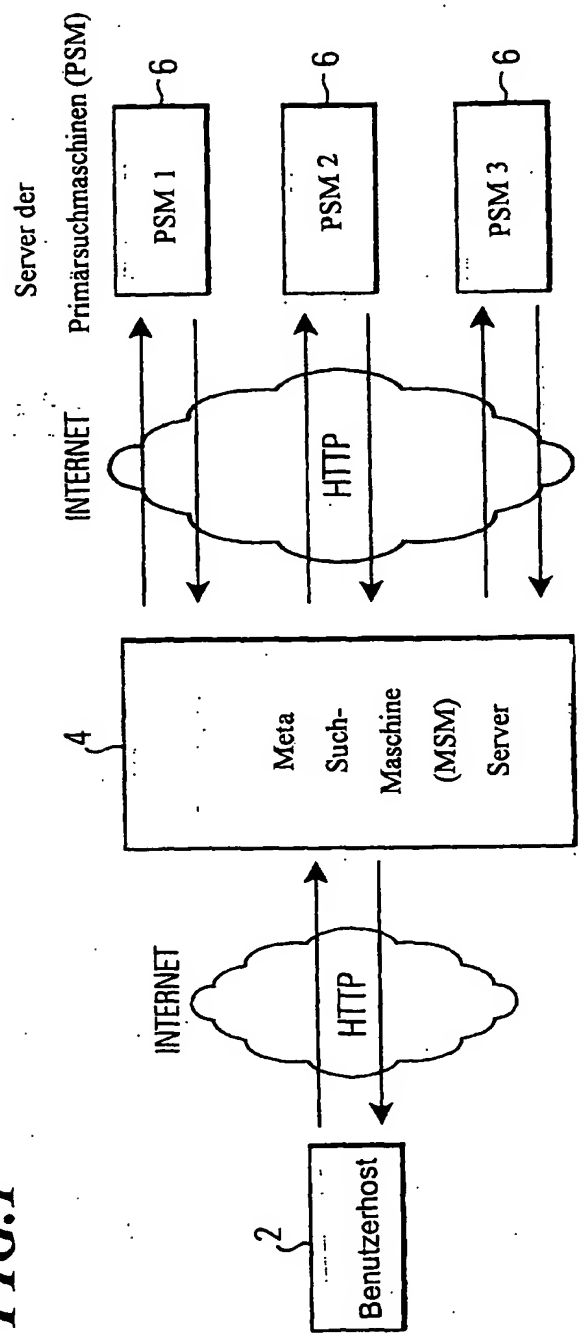
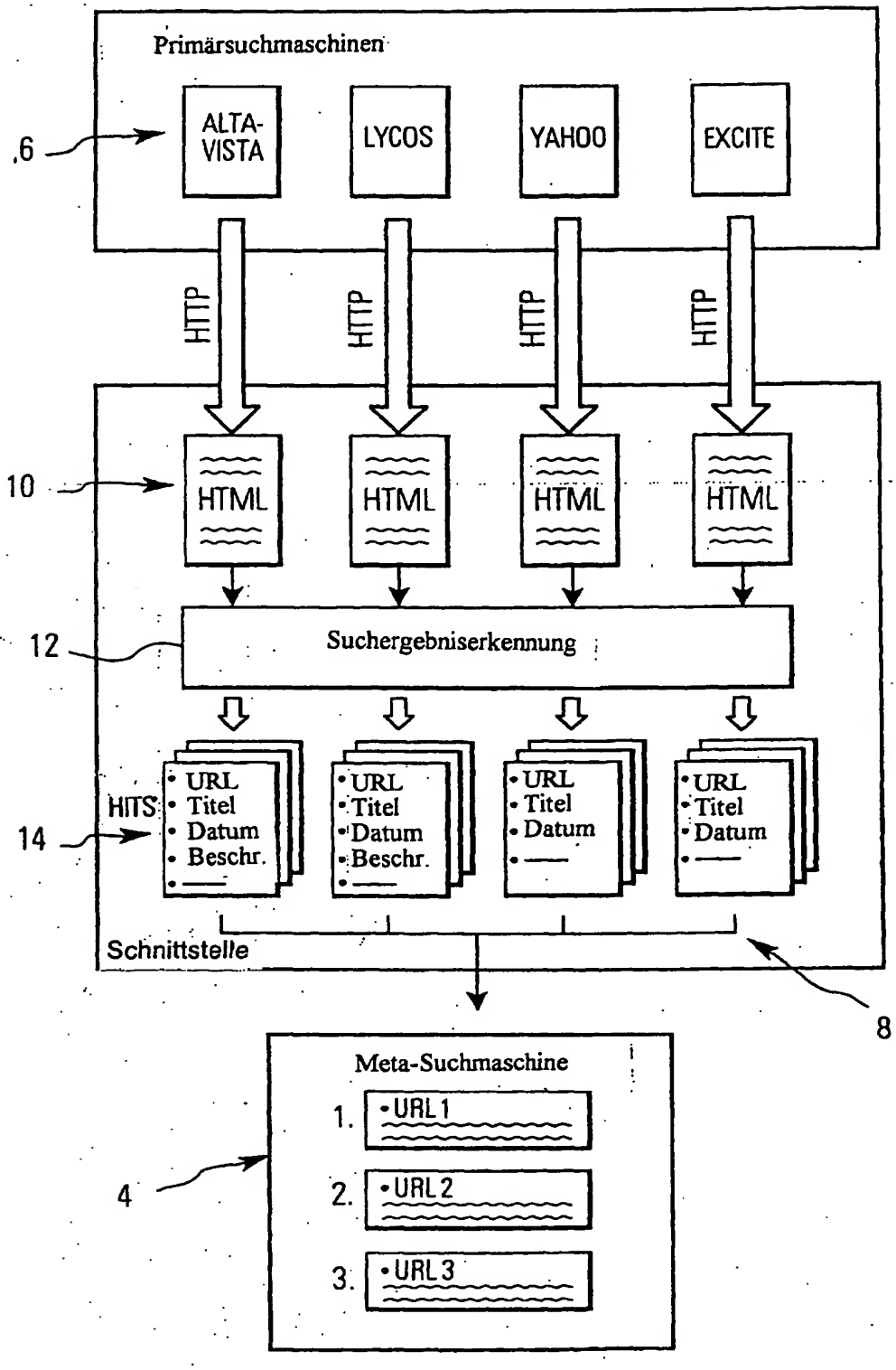


FIG.1

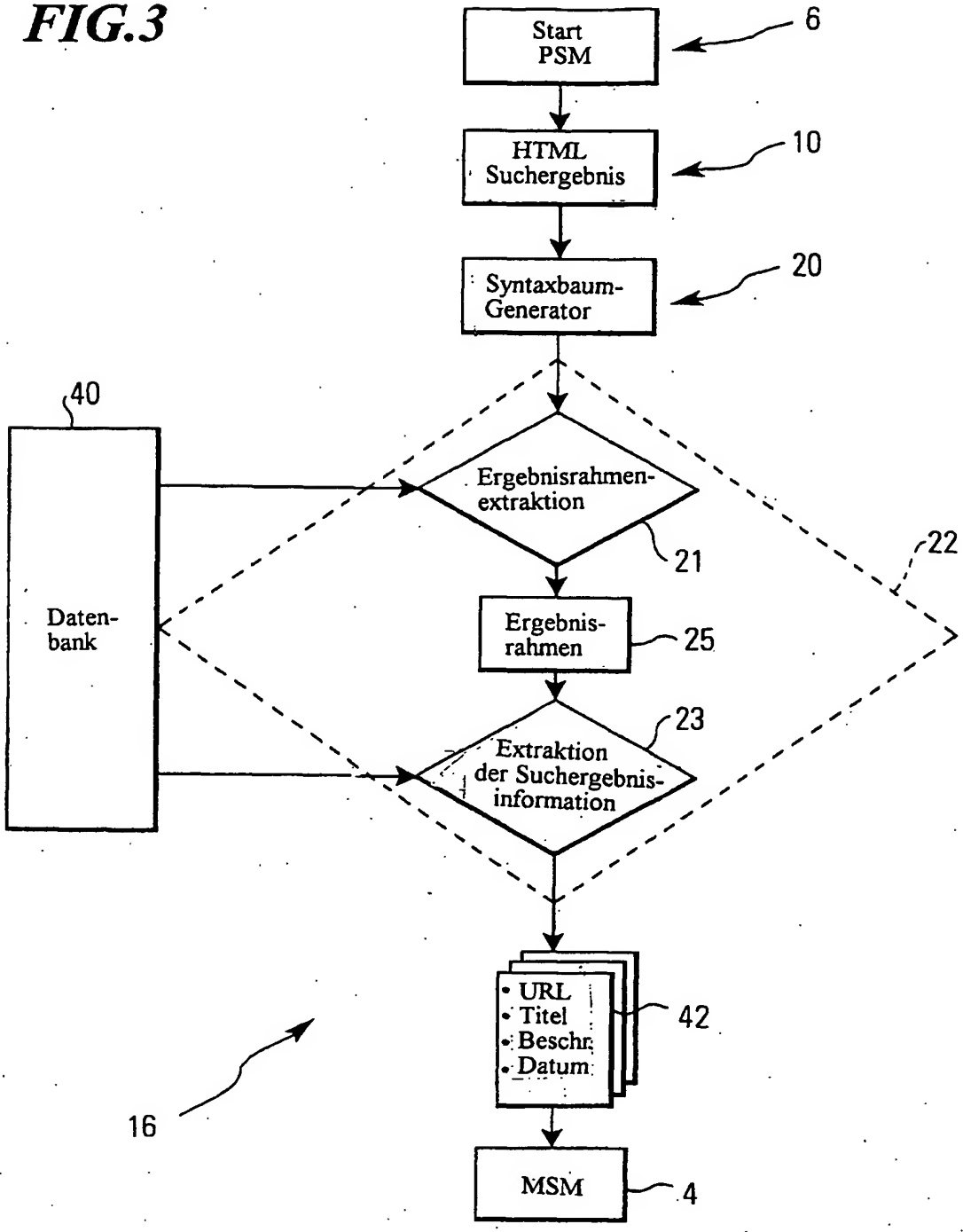
Nummer: DE 100 56 681 A1
Int. Cl.[7]: G 06 F 15/18
Offenlegungstag: 23. Mai 2002

FIG.2



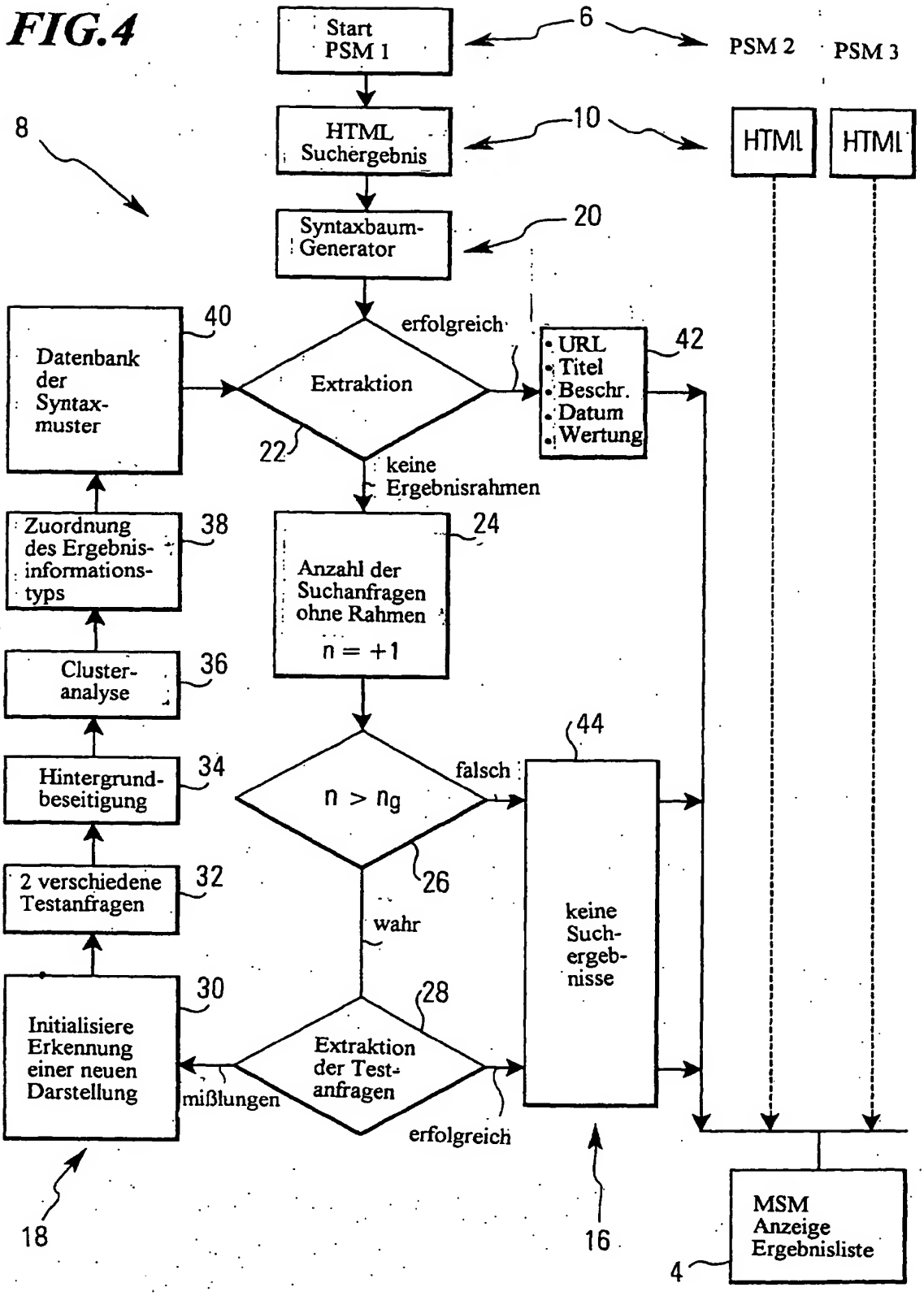
Nummer: DE 100 56 681 A1
Int. Cl.[7]: G 06 F 15/18
Offenlegungstag: 23. Mai 2002

FIG.3



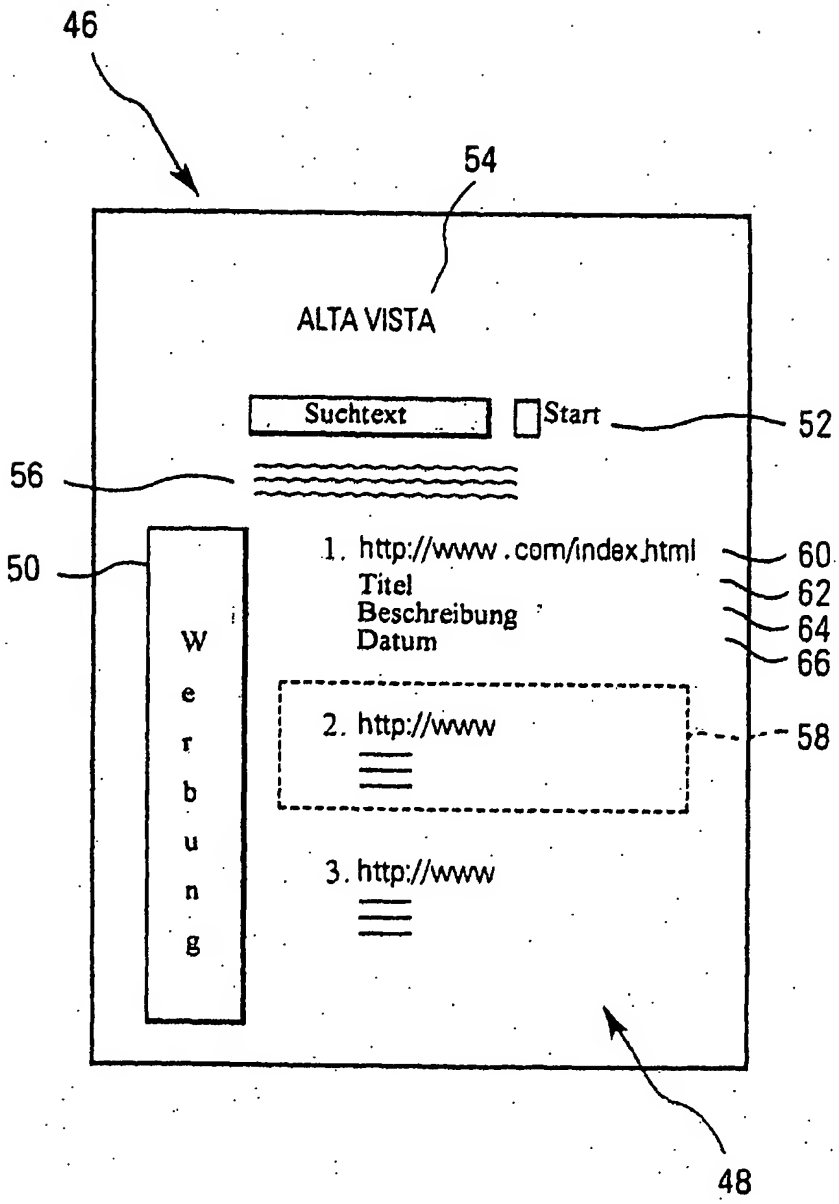
Nummer: DE 100 56 681 A1
Int. Cl.[7]: G 06 F 15/18
Offenlegungstag: 23. Mai 2002

FIG.4



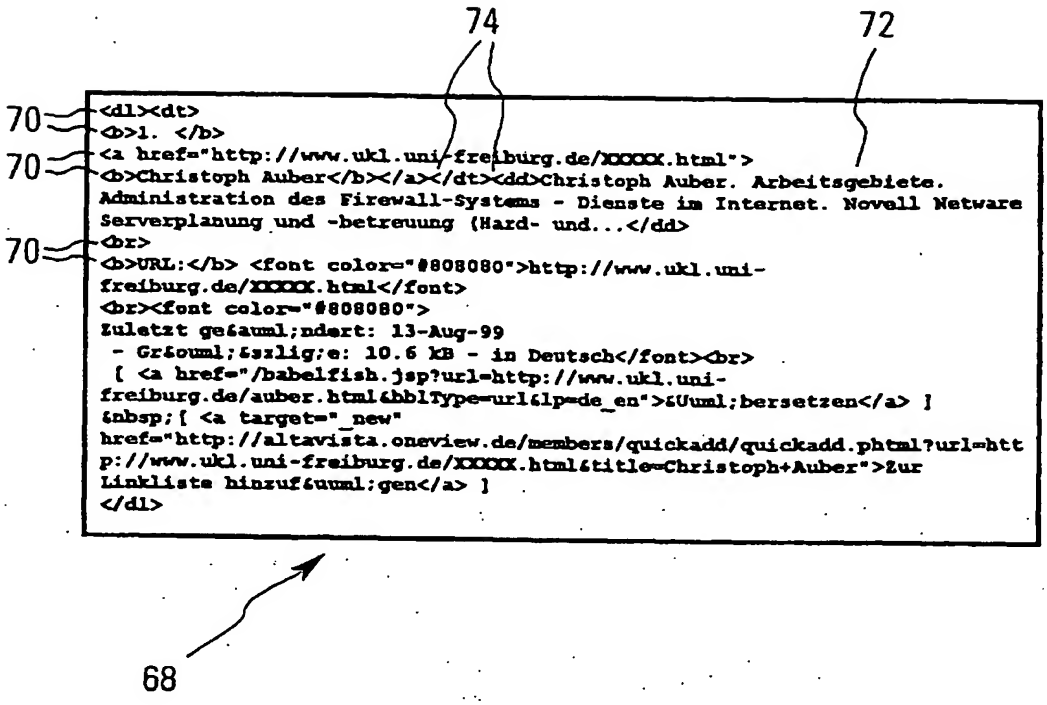
Nummer: DE 100 56 681 A1
Int. Cl.[7]: G 06 F 15/18
Offenlegungstag: 23. Mai 2002

FIG.5



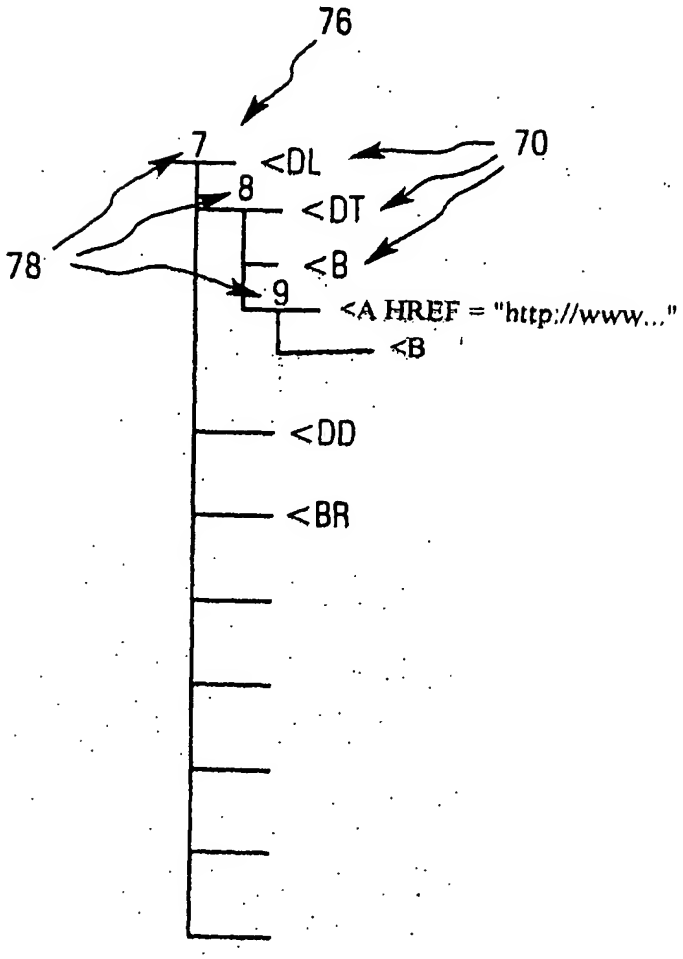
Nummer: DE 100 56 681 A1
Int. Cl.[7]: G 06 F 15/18
Offenlegungstag: 23. Mai 2002

FIG.6



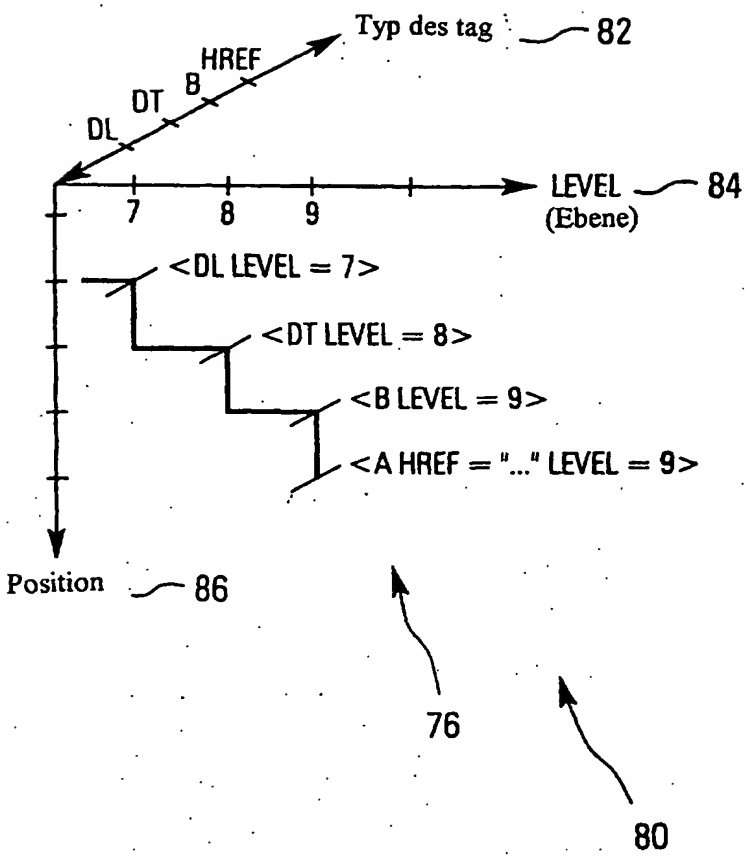
Nummer: DE 100 56 681 A1
Int. Cl.[7]: G 06 F 15/18
Offenlegungstag: 23. Mai 2002

FIG. 7



Nummer: DE 100 56 681 A1
Int. Cl.[7]: G 06 F 15/18
Offenlegungstag: 23. Mai 2002

FIG.8



Nummer: DE 100 56 681 A1
Int. Cl.[7]: G 06 F 15/18
Offenlegungstag: 23. Mai 2002

FIG.9